

Reservoir Sampling

For such situation:

A man has lots of hats. The number is unknown. And he wants to randomly chose one to wear to go to diner. How to design the algorithm?

Each turn, we use a spinner to determine whether we choose this hat or not.

1. For hat #1, the probability to wear it is $1/1$. Because this is the first one and only one in the sample, the probability that we pick it is 1.

2. For hat #2, the probability to wear it is $1/2$.

3. For hat #3, the compatibility probability is $1/3$.

.....

Consider the i th hat, with its compatibility probability of $1/i$. The probability he will be wearing hat i at the time $n > i$ can be demonstrated by a formula.

$$\frac{1}{i} * \left(1 - \frac{1}{i+1}\right) * \left(1 - \frac{1}{i+2}\right) * \dots * \left(1 - \frac{1}{n}\right)$$

For the first term is the probability that i th hat will be accepted.

For the second term is the probability that $(i+1)$ th hat will not be accepted.

For the third term is the probability that $(i+2)$ th hat will not be accepted.

For the last term is the probability that n th hat will not be accepted.

We simplify this formula can get the probability that we still wear i th hat at time n is $1/n$.

In this way, the reservoir sampling algorithm can be used for randomly choosing a sample from a stream of n items, where n is unknown.